Article

Tuomas Vesterinen*

Identifying the Explanatory Domain of the Looping Effect: Congruent and Incongruent Feedback Mechanisms of Interactive Kinds

Winner of the 2020 Essay Competition of the International Social Ontology Society

https://doi.org/10.1515/jso-2020-0015
Published online March 1, 2021

Abstract: Ian Hacking uses the looping effect to describe how classificatory practices in the human sciences interact with the classified people. While arguably this interaction renders the affected human kinds unstable and hence different from natural kinds, realists argue that also some prototypical natural kinds are interactive and human kinds in general are stable enough to support explanations and predictions. I defend a more fine-grained realist interpretation of interactive human kinds by arguing for an explanatory domain account of the looping effect. First, I argue that knowledge of the feedback mechanisms that mediate the looping effect can supplement, and help to identify, the applicability domain over which a kind and its property variations are stably explainable. Second, by applying this account to cross-cultural case studies of psychiatric disorders, I distinguish between congruent feedback mechanisms that explain matches between classifications and kinds, and incongruent feedback mechanisms that explain mismatches. For example, congruent mechanisms maintain Western auditory experiences in schizophrenia, whereas exporting diagnostic labels inflicts incongruence by influencing local experiences. Knowledge of the mechanisms can strengthen explanatory domains, and thereby facilitate classificatory adjustments and possible interventions on psychiatric disorders.

Keywords: looping effect, natural kinds, human kinds, realism, psychiatric disorders, social mechanisms, cultural psychiatry

***Corresponding author: Tuomas Vesterinen**, Department of Philosophy, History and Art Studies/Philosophy, P.O. Box 24 (Unioninkatu 40 A), University of Helsinki, 00014 Helsinki, Finland, E-mail: tuomas.vesterinen@helsinki.fi

1 Introduction

Ian Hacking (e.g. 1986, 1995b) uses the looping effect to characterize a phenomenon that underlies many social constructivist arguments. The looping effect describes the interaction between classifications and the targeted "kinds of people" or human kinds that purportedly share behaviour and traits. The idea is that classificatory practices induce reactions in the members of the human kind by enabling new intentional ways of being and acting. Tracking these changes requires revisions in the original classification, which may in turn lead to further changes in the members of the kind. Consequently, the interaction between the classification and the affected members of the human kind creates a feedback loop that renders the kind a moving target. According to Hacking, this classificatory instability generated by the looping effect distinguishes the human sciences from the natural sciences. In particular, the interactive human kinds studied by the human sciences do not support the robust explanations, predictions and interventions (i.e. epistemic projects) that the natural kinds picked out by the natural sciences do.

Hacking's description of the looping effect has instigated a debate over whether human kinds can be given a realistic interpretation. Cooper (2004) and Khalidi (2010) argue for a realist view on the basis that also some prototypical natural kinds are subject to the looping effect, such as domesticated animals and disease entities. Moreover, Murphy (2006) asserts that looping effects can stabilize human kinds, and Mallon (2016) argues that in general our knowledge can keep up with their rate of change. However, Allen (2018) and Laimann (2018) argue that although some biological kinds are subject to the looping effect, interactive human kinds differ from interactive biological kinds because their classificatory-induced reactions are difficult to explain and predict. Allen associates the problem with classificatory-induced intentional reactions being ontologically anomalous, whereas Laimann associates it with the complex social interactions that underlie our inability to find mechanisms of change and stability. In sum, the looping debate is based on the dichotomy of whether interactive human kinds are real kinds that support robust epistemic projects.[1] Nevertheless, the discussion has mainly concentrated on whether looping effects are problematic for epistemic projects. My aim is to identify how knowledge of looping effects can enhance and supplement explanations of interactive human kinds.

1 I use the term 'real kind' to cover all the kinds that ground robust epistemic projects.

I defend a realist interpretation of interactive human kinds by arguing for an explanatory domain account of the looping effect. I assert that knowledge of the feedback mechanisms that mediate the looping effect can supplement the explanatory domain over which a kind and its variations are accountable. I begin by reviewing Hacking's description of the looping effect and the discussion it has instigated. I then argue that intentionally mediated looping effects are causally explanatory and hence the ontology of human kinds is irrelevant to my explanatory account. Next, based on the contrastive-counterfactual theory of explanation (Woodward 2003; Ylikoski 2007), as well as on the property cluster view of natural kinds (Boyd 1999), I assert that mechanistic explanations of human kinds have *applicability domains* over which they stably account for the kinds' properties under counterfactual situations. The applicability domain is better when it holds over more properties of the kind in a wider range of alternative situations. This means that explanations of a homeostatic property cluster kind fall on a continuum of goodness described by the applicability domain. The idea is that a better mechanistic explanation of a kind enables more secure domain-relative projections (i.e. generalizations and predictions) based on the kind. Moreover, an explanation ideally spells out its applicability domain because human kinds support limited epistemic projects. In that case, classificatory projects that apply the explanation do not exceed the limits of their own applicability. Based on this account, I argue that knowledge of feedback mechanisms can supplement the domain over which a human kind is explainable by accounting for some of its dynamic properties and identifying the limits of the domain's applicability. Nevertheless, the actual explanatory relevance of a feedback mechanism is an empirical question because interactive human kinds are affected in different ways and to different degrees by looping effects.

In the last section, by applying the explanatory domain account to cross-cultural case studies of psychiatric disorders, I argue that there are two types of feedback mechanisms that mediate the looping effect. Congruent feedback mechanisms describe matches between classifications and kinds. They can supplement explaining, for instance, the cross-cultural variation of schizophrenia. On the other hand, incongruent feedback mechanisms describe mismatches between classifications and kinds. They can help to identify unintentional misclassifications and their effects, or more problematically for epistemic projects, value-driven classificatory and behaviour adjustments. For example, classificatory practices that disregard culture-dependent manifestations of disorders exceed their applicability and hence cause incongruence. Finally, I suggest that knowledge of the feedback mechanisms facilitates classificatory adjustments and interventions on interactive kinds such as psychiatric disorders.

2 The Looping Effect and Interactive Kinds

2.1 The Looping Effect, Interactive Kinds and Realism

The looping effect's plausibility as a demarcation criterion between the natural and human sciences depends on the nature of the generated instability and whether it prevents interactive kinds from being real kinds. As examples of the kinds subject to the looping effect, Hacking (1995b, p. 351–352) has concentrated on "kinds of people" defined by their behaviour, condition, actions, tendencies, emotion and experience. For example, he has provided case studies of psychiatric disorders, such as fugue, multiple personality disorder, schizophrenia and autism spectrum, in addition to cases like teenage pregnancy, child abuse and homosexuality (see Hacking 1986, 1995a, 1998, 1999). According to the analyses, classificatory activities have affected the targeted kinds to such a degree, that the classifications have had to be amended.

Hacking has provided a particularly detailed analysis of the looping effect of multiple personality disorder (Hacking 1995a). Until the end of the 1970s, the diagnosis was based on individuals having two or three alternating personalities. However, once knowledge about the syndrome spread among specialists, media and lay people, the number of diagnoses as well as the diversity and number of alters associated with the syndrome started to increase. Hacking argues that popular knowledge about the syndrome created a prototype, which in turn induced more people to conform with it or otherwise react to it in varied ways. These reactions needed to be explained and to be integrated into the classification, which again affected the syndrome. Consequently, by the 1990s people were diagnosed with having hundreds of fragmented alters.

The reactions generated by the looping effect should be distinguished from at least two other ways in which kinds can change. First, the kind may change within the range set by the classification for non-classificatory reasons, and therefore there is no need to amend the classification. Second, the kind may change outside the range set by the classification for non-classificatory reasons. In such a case, the classification may need to be amended to match the changes. However, I am solely concerned with the third type of case, where the targeted kind is affected by the classification and may change because of it.

Although Hacking's examples mostly concern scientific classificatory practices and their objects, folk categories and their objects are subject to the looping effect as well. In general, scientific classifications reinforce and sharpen the boundaries of pre-existing folk categories or create new categories that become folk categories (see Root 2000, p. 631). An example of the former is

psychiatric disorder terms. Madness was historically an unspecified folk category that has become increasingly sharpened and recategorized due to scientific research, thereby affecting the behaviour of the classified people. Hacking (1986, 2002) describes homosexuality as an example of the latter. Following Foucault (1978), Hacking argues that the homosexual as a kind of person only came into being through legal and medical thinking. Subsequently, the kind of person was transformed when individuals began to adjust their self-descriptions and actions in reaction to the classification and emergent prototypes. Finally, the 'gay movement' changed the values and beliefs associated with the classification.

A classification may either generate merely reactions from the classified or also genuinely shape the attributes targeted by the classification. In the first case, a classification may prompt reactions due to the status or stigma associated with it, while the targeted kind-typical behaviour is not affected. According to Hacking (1999, p. 114), for instance, once auditory hallucinations became an integral part of schizophrenia diagnosis, people ceased to report them. This in turn led to their diminished importance in the diagnosis. In the more substantive feedback effects, classifications and associated stereotypes influence the behaviour and content of the kinds targeted by the classification. Along these lines, Hacking (1999, p. 114) mentions that the content of the hallucinations also changed due to becoming diagnostically important. In such cases, the classification does not merely successfully or unsuccessfully pick out a pre-existing kind, but is part of the casual structure of the social world, and may prompt reactions that lead to changes in the kind as well as the classification itself. If the classification is perceived as a match and meaningful to the classified and the larger social audience, it may generate congruence (i.e. conformity with the classification), while a perceived mismatch may lead to capricious behaviour and increased incongruence (i.e. nonconformity) (see Table 1 and Section 3.2).

Classifications of human kinds induce reactions primarily because they are value-laden (Hacking 1995b: 370). Although Hacking concentrates mostly on negatively value-laden classifications, positively viewed human groupings, especially ones that have political importance, are also subject to the looping

Table 1: Looping effects.

	Status reaction	**Kind looping**
Perceived match	No modification	Congruence
Perceived mismatch	Classificatory modification	Incongruence

effect.[2] This is exemplified by how the idea of an aristocrat as a kind of person was connected with high expectations concerning character and behaviour. The expression *noblesse oblige*, for example, describes the conferred social obligations that were associated with aristocratic titles. These perceived obligations interacted with the behavioural patterns and traits of the aristocrats.

The looping effect seems to challenge classificatory (or kind) realism in the human sciences. In general, realism about natural kinds can be approached from an essentialist or naturalist perspective (Bird and Tobin 2008; Kornblith 1993; Reydon 2009). The essentialist approach defines kinds as natural, based on shared necessary and sufficient conditions that are determined by underlying intrinsic properties (Ellis 2001; Kripke 1980; Putnam 1975). In addition, according to this approach, natural kinds are usually thought to be upheld by laws of nature. On the other hand, a weaker form of realism defends natural kinds from an epistemic and naturalistic point of view without committing itself to essentialism. It stresses how natural kinds ground inductive inferences, explanations and predictions (Boyd 1999; Dupré 1993; Millikan 1999). In short, according to naturalism, natural kinds are needed to explain why some scientific classifications ground epistemic projects more than others. According to Hacking, however, neither approach is applicable to the kinds targeted by the looping effect.

Hacking (1983, 1991, 2007a) maintains that there are no prepackaged kinds united by their common naturalness, but instead there are different ways of classifying that correspond with the varying nature of the targeted kinds. Some natural kinds classified by the natural sciences are better described according to the essentialist approach, while others according to the naturalist tradition (Hacking 1991, p. 123).[3] But more importantly, whereas the natural kinds studied in the natural sciences are indifferent to our classifications and manipulations, classificatory activities in the human sciences lead to the looping effect that renders the targeted human kinds moving targets or interactive kinds (Hacking 1999). The reason is that social phenomena do not constrain classificatory possibilities to the same extent as natural phenomena (Martínez 2009, p. 214). Because of the dynamic nature of the kinds studied by the human sciences, Hacking has labelled his view as dynamic nominalism. In short, he uses the looping effect to draw a principal distinction between the instability of interactive kinds and the stability of natural or indifferent kinds.

The argument from the looping effect to non-realness (or nonnaturalness) of interactive kinds is not straightforward. Clearly, because the looping effect may

2 Hacking (1997, 2007a) mentions "genius" for romantics as an example.

3 Lately Hacking (2007b) has argued that the concept of natural kind has become obsolete because of the confusion it creates.

lead to property changes, interactive kinds cannot be essential kinds. However, it is not clear why the looping effect would preclude all interactive kinds from grounding robust epistemic projects. Hence, there is an ontological and an epistemic question that require clarification: what renders interactive kinds susceptible to the looping effect, and why would their instability be epistemically problematic for classifications? The questions are co-dependent, but it is analytically helpful to keep them separate.

2.2 The Looping Debate

Hacking's account of the looping effect has stirred a debate over whether the affected kinds can be given a realistic interpretation. Critics argue that the looping effect does not preclude human kinds from being real kinds because some prototypical natural kinds are interactive and some looping effects of human kinds are mediated by environmental changes. For example, Douglas (1986, p. 101) has pointed out that microbes adjust themselves to our classificatory and medical interventions. Especially Cooper (2004) andKhalidi (2010, 2013) have developed this idea by arguing that the looping effect is not restricted to the examples involving human behaviour that Hacking offers, but affects biological kinds as well. For instance, labelling some species as domestic animals has led to selective breeding, which in turn has eventually created new breeds of dogs and cats, for example. Consequently, we have had to adopt new labels and taxonomies to match these changes. Similarly, labelling bacteria and viruses as diseases leads to medical interventions that change them through natural selection. This in turn requires adjustments in treatments and labels. The gist of Khalidi's and Cooper's argument is that since some biological kinds clearly ground inductive inferences and explanations, there is no reason why other interactive kinds cannot do the same.

Hacking (1999, p. 106), however, has pointed out that genuine looping effects are mediated by the awareness of being classified, thereby distinguishing interactive kinds ontologically from indifferent kinds. This interpretation is supported by Hacking's (1986, 1995a, ch. 7) relying on Ascombe's theory of intentional action under a description. Once a new description, associated with a classification, becomes socially available, it prompts reactions by enabling new conceptual possibilities for being and acting. These new actions, in turn, render the kind a moving target. Cooper (2004) and Khalidi (2013), nevertheless, argue that even according to Hacking himself humans need not always be aware of the classification for it to influence their kind-typical behaviour. For instance, a child labelled as having attention deficit hyperactive disorder (ADHD) may be

placed in a school where "stimulant-free schoolrooms" influence her behaviour (Hacking 1999, p. 103). Similarly, an individual can acquire refugee characteristics by being part of a refugee group (Hacking 1999, p. 32). Furthermore, realists argue that most of the time the instigated changes in human kinds are not fast enough to rule out epistemic projects (Mallon 2016), and looping effects can be stabilizing as well as destabilizing (Griffiths 1997; Kuorikoski and Pöyhönen 2012; Murphy 2006).

However, Laimann (2018) and Allen (2018) have argued that interactive human kinds differ from non-human kinds because they are prone to wayward behaviour. Laimann (2018) argues that looping effects render human kinds epistemically capricious because their behaviour often invalidates existing classifications and knowledge about them. She associates the problem with the complex nature of human interactions that undermine our ability to discover mechanisms that underlie patterns of change and stability. Allen (2018) comes to the same conclusion by arguing that intentional action is ontologically different because it enables humans to fake or be mistaken over their kind membership. Arguably, these creative reactions are harder to predict and rectify than mistakes over non-human (or non-aware) classifications.

These problems are not restricted to feedback effects that are mediated by subjective awareness on the part of the classified individuals. For example, apparently one of the reasons for the ADHD epidemic in the USA is that some schools and caretakers pressure clinicians for diagnoses to obtain medicine and improved learning facilities (e.g. tutoring, smaller classes) for children (First 2017). This makes misdiagnoses probable, leading to general diagnostic distortions. These problematic feedback loops can also occur when children act under the description of ADHD, without being explicitly aware of the classificatory description. It is enough that they learn the characteristic action pattern from the people in their "in-group". In fact, it is probable that a person's symptom profile depends on the specifics of the mediating feedback mechanisms and their complex interactions with the underlying psychological and neurocognitive processes, as well as on the larger social and cultural forces.

In conclusion, the creative and complex nature of human reactions to being classified complicates our ability to explain and predict kind-typical behaviour. However, in the next section, I argue that interactive human kinds, and the intentional actions that generate them, do not require a different type of explanation from non-intentional explanations. Therefore, the question about the ontological difference of interactive human kinds can be set aside. In the third section, I defend an explanatory account of feedback effects according to which

their complexity does not preclude interactive human kinds from sustaining scientifically relevant epistemic projects.

2.3 Explanation and Epistemic Instability

The looping effect as a demarcation thesis can either be interpreted as a strong claim in favour of anti-naturalism about explanation, or as a weaker argument about causal construction which is compatible with naturalism. The former would mean that the ontological nature of human kinds in principle precludes them from being real kinds, whereas according to the latter approach the problem is not principal, but instead an epistemic problem due to causal complexity.

As an anti-naturalistic argument, the looping effect would be a thesis in favour of non-naturalist interpretivism, and would support the separation between the kinds studied by the human and natural sciences. Interpretivists generally argue that humans are self-interpreting, and therefore understanding intentional action requires interpreting its meaning to the agent, instead of explaining it by causes or laws (e.g. Geertz 1973; Taylor 1971; Winch 1958). Moreover, arguably interpretations need to account for the culturally situated and holistically determined beliefs, concepts, categories and the like. These make actions meaningful to agents and understandable, in the light of their context, to observers. Hence, one could hold that the efforts to explain human groups with causal generalizations induce new self-interpretations whereas the aim should be to understand the humans according to their own meanings and concepts. The problem is that this muddles the distinction between classifications and kinds. One way to understand the problem is that interpretivism seems to set the produced action in a conceptual (or quasi-logical) connection with its reasons (von Wright 1971). That is, while actions are identified based on the agent's own reasons (i.e. beliefs and desires), those reasons can only be established based on the actions they are reasons for (see Rosenberg 2016, ch. 3). Consequently, behaviour could be interpreted and conceptualized only in retrospect, making objective and generalizable classifications virtually impossible. Lastly, the conceptual connection could describe how classifications constitute human kinds, not only how they can be explained. Nevertheless, in all these cases, generalizations of human kinds would be mostly exhausted by their classificatory descriptions, and hence would not be real kinds that ground robust projections.

In the following, I will first argue against the non-naturalist explanatory view of the looping effect, and thereafter against the stronger constitutive view. The general view in the philosophy of social sciences is that reasons can function as causes in explanations and that interpretation and causal explanation need not

be mutually exclusive (Henderson 1993; Kincaid 1996; Tuomela 1977; Ylikoski 2001). As an example, when an anthropologist conducts fieldwork by interpreting local customs and behaviours, she relies on the causal efficacy of cultural structures and beliefs. Especially the contrastive-counterfactual theory of explanation (Ylikoski 2001; Woodward 2000, 2015) matches the thesis that explanations in the social and natural sciences do not differ in key features. According to the theory, explanations provide descriptions of objective causal (and constitutional) relations by describing counterfactual dependencies that answer *what-if-things-had-been-different* questions (*what if*-question). Explanations, in addition, have a contrastive structure, so that they answer questions of why fact rather than foil. The contrastive structure makes explicit the aspects to be explained and thereby determines whether a putative explanation is relevant. Hence, reasons are causally explanatory because they can provide counterfactual information on why someone acted in one way rather than another. This means that a putative intentional explanation is explanatory if it can answer questions concerning how the explanandum action would have been different, had the relevant beliefs and desires been different (see Ylikoski 2001, p. 97). In this light, a putative feedback explanation is explanatory if it can describe how a difference in classificatory related conceptions and beliefs would have made a difference to the behaviour of the classified people. But having said that, understanding the causal process may require resorting also to lower-level explanations as well as structural explanations.

The constitutive view of human kinds can be understood as a form of conventionalism (cf. Kornblith 1993). A strong interpretation of conventionalism would mean that human kinds are merely subjective distinctions or groupings made by scientists qua scientists. However, this would trivialize human kinds to the extent that kind membership would be merely a matter of opinion or opinions (see Griffiths 1997, p. 198). A weaker conventionalism would mean that interactive human kinds are akin to institutional facts which, according to Searle (1996), are epistemically objective although constituted by collective acceptance. Crucially, however, interactive human kinds are not constituted merely by collective representations, or by representations in the minds of scientists qua scientists, but also by the behavioural patterns and traits they bring about. Thus, the conceptions and beliefs associated with the label multiple personality disorder do not constitute multiples as kinds of people, but those beliefs as part of classificatory practices (treatments, institutional infrastructures, etc.) and prototypical expectations, may constrain, shape and enable kind-typical behaviours and traits. This is exemplified in how a posteriori research is needed to uncover the reasons that brought about the behavioural pattern or its reinforcement. And since I already argued that reasons can function as causes in kind explanations, nothing in principle prevents

looping effects from describing causal interactions between classificatory descriptions and human kinds.

Shared conceptions may nonetheless be necessary to enable the complex social interactive processes that bring about and sustain human kinds. However, it is not plausible that new human kinds and kind-typical actions are born simply from conceptual stipulations. Instead, it seems credible that there are different degrees of conceptually and socially induced kind-typical intentional actions. This means that novel kinds come about incrementally, so that the kind and its conception are egging one another on in various ways. As an example, the diagnostic category of ADHD has grown more specific through institutional and social interactions (Lakoff 2000). In addition, kind-typical intentional actions may become possible before their linguistic expressions. The idea is that feedback processes of social interactions can brings about human kinds by generating compelling emotional experiences (see Collins 2004, p. xii). For instance, Siegel (1997) argues that Indonesian national identity first guided behaviour rather as a structure of feeling than as an explicit category.

The antirealist view of interactive human kinds can also be defended naturalistically so that the looping effect is a form of causal construction that destabilizes the affected kinds (see Hacking 1995b, p. 362). However, those who take a realist approach to human kinds argue that looping effects can be stabilizing and that in general our theoretical knowledge can keep up with their rate of change (Mallon 2016; cf. Murphy 2001). Laimann (2018) claims, instead, that interactive human kinds are generally capricious and problematic for epistemic projects because of the difficulty in discovering the mechanisms that underlie their patterns of change and stability. The reason is that feedback mechanisms can interact in complex and unpredictable ways with each other, and with larger social circumstances. Consequently, she argues that secure extrapolations based on kinds in the human sciences are difficult if not impossible. This interpretation would mean that human kinds are historical in the sense that we can explain their alterations and context dependent stability only in retrospect.

I agree with Laimann's general idea to the extent that our ability to explain human kinds and their property variations, rather than the superficial stability of the kinds by itself, is the key to measuring the epistemic stability and scientific relevance of human kinds. This means that knowing how the kind would vary under different circumstances supports its realist interpretation. However, since I have argued that there is nothing in principle preventing human kinds from supporting epistemic projects, whether they do support projects is an empirical question. Next, I argue that the looping effect's explanatory relevance, and the epistemic projects that the human kinds support, are not all or nothing matters.

3 An Explanatory Account of the Looping Effect

3.1 Explanatory Domain of Feedback Mechanisms

In this section, I argue for an explanatory domain approach to the looping effect. A common realist approach to interactive human kinds has been to interpret them as homeostatic property clusters (HPC view) (Griffiths 1997; Hauswald 2016; Kokkonen and Koskinen 2016; Kuorikoski and Pöyhönen 2012; Pöyhönen 2013). According to the HPC view (Boyd 1999), natural kinds consist of homeostatic property clusters that are reliably generated by causal mechanisms. The members of a kind may share different properties and the members as well as the kind may change over time. The HPC view can accommodate social mechanisms as external mechanisms that maintain property clusters (Boyd 1991; Mallon 2003). Moreover, it is commonly argued that uncovering social mechanisms enables extrapolations (Elster 2015; cf. Steel 2008).[4] According to Kuorikoski and Pöyhönen (2012, p. 191), the reason is that identifying underlying mechanisms, not merely patterns and regularities, enables inferences to alternative situations. They point out that this matches the HPC view's idea that while causal mechanisms explain the clustering of properties, the clusters together with their mechanistic explanations enable secure extrapolations and projections (see also Reydon 2009). Therefore, the more we learn about the mechanisms that generate property clusters, the more securely we can extrapolate that token clusters are of the same type. Moreover, when a specimen is identified as a member of a well explained kind, we can reliably generalize and predict its behaviour based on that membership.

My account is that the looping effect is mediated in some cases by feedback mechanisms that can supplement explaining the clustering of properties of human kinds. A model of the mechanism(s) that generates a property cluster human kind ideally specifies a *domain of applicability* over which it stably explains the kind and its property variations. The domain of applicability has a *scope* and a *depth* dimension (cf. Griffiths 1999, p. 217). The *scope* dimension describes the actual properties, places, and times where the explanation is applicable. For instance, an explanation can cover some properties of a human kind for a determined duration and location. The *depth* dimension describes the counterfactual stability of the explanatory relation, that is, how dependent the inferences that the explanation enables are on non-included situations. My argument is that explanations of a homeostatic property cluster fall on a continuum of explanatory power or goodness described by the domain of applicability and the contrastive-counterfactual

4 Roughly, a social mechanism can be said to consist of individual agents whose relations and actions are responsible for a social phenomenon, see Hedström and Ylikoski (2010).

theory. Nevertheless, a mechanistic model of a kind is explanatorily relevant only if it can systematically account for some of the kind's properties under counterfactual situations. This would mean that the mechanistic generalization is stable (i.e. invariant) under hypothetical interventions (Woodward 2000). Crucially, applicability domains are limited in the human sciences because social phenomena are contingent on complex circumstances. Thus, an explanation can exceed the limits of its spatial and temporal applicability by disregarding the dynamic nature of the targeted kind. An explanation should therefore make explicit and explain why the given applicability domain is optimal for the set of phenomena. This is tantamount to identifying better the limits of the underlying causal mechanistic structure of the phenomena. Hence, if feedback mechanisms are part of the causal structure of human kinds, knowledge about them can support more secure domain-relative extrapolations and projections.

The explanatory domain of applicability can be illustrated with Luhrmann et al.'s (2015) study on the nature of the hallucinatory voices schizophrenia patients hear in the USA, Ghana and India. The study indicates that the voice-hearing experiences are exceptionally harsh in the USA in comparison to Ghana and India. The study seems to suggest that one of the reasons for the harsh voices in the USA is a feedback mechanism between the prototypical expectations associated with the diagnostic category and the hallucinatory voices. In this light, the feedback explanation's scope is the nature of the hallucinatory voices schizophrenia patients hear in the USA during a specified time-scale, whereas the explanatory relation is counterfactually relatively stable given that the diagnostic expectations are institutionally and socially entrenched. This means that if we or some social process were to manipulate the diagnostic expectations during that time, the nature of the voice-hearing experiences would change. Finally, such an explanation should spell out why its scope is limited to the USA and whether it has exceptions. Moreover, it should be established that the explanatory depth is correct for the explanation's scope, so that, for example, lower-level details, such as schizophrenia's different genetic subtypes, are irrelevant to explaining the voices in the USA in contrast to India and Ghana.

The explanatory domain account offers a method for identifying how relevant a feedback effect is in explaining a human kind (cf. Pöyhönen 2010, 2014). A stable feedback explanation can supplement the domain of applicability of a kind explanation. Alternatively, knowledge of a feedback effect may help to identify how a putative explanation is unstable. Furthermore, feedback explanations fall on horizontal and vertical axes that represent their ability to supplement the applicability domain. An explanation is enhanced horizontally by widening its scope, or vertically by providing a deeper explanation within the scope. Although it is commonly argued that wider explanations are less deep, I rely in my

explanatory analysis mostly on strong complementarity (Marchionni 2008), so that relevant feedback explanations can enhance both dimensions. This strategy is important for explaining interactive kinds that have inseparable biological, psychological and social properties. In such cases, integrating higher-level explanations, such as feedback mechanisms, with lower-level ones, can strengthen the explanation's applicability domain. In the schizophrenia case, the feedback mechanism can both provide a deeper explanation of the disorder's core symptoms in the USA, as well as account for why the explanation's scope is limited to the USA.

The depth dimension of a feedback mechanism's explanatory domain can be further explicated with the contrastive-counterfactual theory. The depth of an explanation depends on how many relevant *what if*-questions it can answer. Thus, a better explanation of a property cluster kind answers more counterfactual questions because it locates the kind within a larger space of alternative possibilities. A feedback explanation can contribute to the explanation by providing fine-grained information of the counterfactual dependence between classificatory descriptions and properties of the kind. In this case, the feedback generalization describes how the kind would change, if its classification were to change in diverse ways, and vice versa. This counterfactual dependence can be explicated with explanatory *insensitivity* and *precision*.[5] The insensitivity of an explanation describes the invariance of the explanation under different background conditions (Woodward 2000; Ylikoski and Kuorikoski 2010). In other words, if a feedback mechanism makes a difference to the targeted aspects of the human kind (i.e. the explanandum), including it into explaining the kind (instead of leaving it as a background condition) would enable the explanation to answer more *what if*-questions. An explanation of schizophrenia that includes a feedback explanation of diagnostic expectations is more insensitive if the expectations make a difference to the psychiatric disorder. The explanation's precision describes its ability to characterize in a fine-grained way why something is the case in contrast to something else (Ylikoski and Kuorikoski 2010). In the schizophrenia case, the feedback explanation could make an explanation of the psychiatric disorder more precise by describing in (more) detail the nature of schizophrenia voices in the USA in contrast to their nature in India and Ghana, or better still, everywhere.

The applicability domains of feedback explanations can also be roughly compared. For instance, since severe autism has a strong neurobiological basis, and the individual's ability to communicate is limited, looping effects are not mediated by intentional reactions. The looping effect may nonetheless explain

5 Other explanatory virtues could be relevant as well, see Ylikoski and Kuorikoski (2010).

some behavioural responses to the actions of caretakers and environmental changes (Kuorikoski and Pöyhönen 2012). Moreover, a classificatory feedback explanation of pathogenic bacteria can supplement evolutionary explanations to account for antibiotic resistance. Nevertheless, the feedback explanation alone is unable to explain how subtle differences in classification would have made a difference to the bacteria. In sum, these feedback explanations are not as stable as the feedback explanation of schizophrenia.

Classificatory feedback mechanisms can be compared with self-fulfilling prophecies. They are based on expectations becoming a key component of the causal mechanism that generates the expected outcome (see Biggs 2009). This means that the prophecy can be invalidated if relevant people learn to intervene on the mediating causal mechanism. Consequently, if classificatory feedback mechanisms are self-fulfilling prophecies, disseminating knowledge about them could diffuse their causal efficacy. However, a classificatory feedback explanation's causal power to diffuse itself is limited. As Mallon (2016) argues, many social categories are firmly entrenched in larger social and material environments, thereby restricting one's space for action even if one becomes aware of their social nature. This is one of the reasons why true social change requires a highly concerted effort. Conversely, not just any enforced expectation or arbitrary claim will initiate a kind shaping or enabling looping effect. The reason is that feedback mechanisms bring about or reinforce interactive properties of kinds by interacting with other factors. It is commonly agreed, for example, that psychiatric disorders need multifactorial explanations that combine social, psychological and biological causal mechanisms and causes (Kendler, Zachar and Craver 2011).

The modularity requirement for mechanistic models provides a way to understand the limits of feedback explanations. According to Woodward (2002), a model is modular if an intervention on a putative cause does not alter the subsequent causal relations in the underlying structure of the represented causal mechanism. The unaltered causal structure ensures that the model can predict the outcome of an intervention. Building on this idea, Steel (2006) argues that some interventions in the human sciences violate modularity because they are structure altering.[6] This means that the interventions cause unpredictable changes in the causal relations between the parts comprising the modelled social mechanism. In this light, the looping effect describes how classifications as part of classificatory or bureaucratic practices (understood as interventions here) alter the structure of interactive kinds inadvertently. The reason can either be epistemic, as argued here, so that sometimes the changes can be explained and anticipated by unboxing the mediating feedback mechanisms and their interactions with other mechanisms.

6 Such critiques have been influential in economics, see Lucas (1976) and MacKenzie (2008).

On the other hand, antirealists seem to claim either that the interactions are too complex to be modular, or that human reactivity eschews modularity in principle. This would also mean that if a putative feedback explanation is incorporated in classificatory practices, it will become less explanatory by breaching the original feedback structure. However, whether a feedback structure can be modelled so that it enables some interventions, is ultimately an empirical question. Indeed, a model need not provide a complete description of a feedback system, and its interaction with other mechanisms, to enable domain-relative interventions and causal predictions. Next, I provide some empirical evidence that knowledge of the feedback mechanisms' applicability domains facilitates explanations and predictions of interactive kinds as well as interventions on them.

3.2 Congruent and Incongruent Feedback Mechanisms

In the following, I employ empirical case studies to argue that the looping effect is mediated by congruent and incongruent feedback mechanisms. They are abstract and rough models that need to be filled with empirical details to generate generalizations and predictions. Understanding these feedback mechanisms is especially relevant for identifying and explaining psychiatric disorders. Knowledge of the mechanisms can supplement, and help to identify the limits of, the applicability domain over which a disorder is stably explainable. This knowledge can also facilitate mitigating negative feedback effects.

Congruent mechanisms explain how feedbacks generate, reinforce and maintain stabilizing loops between classifications and interactive kinds. Intentionally mediated congruence ensues when the classification is found meaningful and natural by the classified so that, for example, it seems to explain and exonerate one's experience, condition and behaviour (see Hacking 1995b).[7] According to Mallon (2016, p. 73–93), classified behaviour as a social role can also be preferred for strategical reasons, reinforced culturally and amplified by non-intentional automatic processes. He maintains that these causal mechanisms may lead, under the right circumstances, to a social role becoming structurally entrenched in social, material and institutional environments. The idea is that structurally entrenched social roles are stable enough to be homeostatic property cluster kinds. I interpret this so that the interaction between the mentioned psychological and social mechanisms (and other factors) may form a feedback mechanism that explains congruence between classifications and kinds.[8] In this case, the

7 See also Appiah (2005) and Haslanger (2012).

8 The interaction between the mechanisms can be understood in the light of structural individualism, see Coleman (1990) and Hedström and Ylikoski (2010).

classificatory practices and conceptions do not only provide opportunities and constrain behaviours from above but also render the behaviours and experiences meaningful and seemingly natural. However, the extent to which interactive human kinds support scientifically relevant epistemic projects does not only depend on the stability of the kinds, but also on our ability to explain their properties under domain-relative alternative circumstances. Consequently, it is crucial to identify how congruent mechanisms can supplement the domain over which kinds are explainable.

Congruent mechanisms described by the labelling theory can explain why classified members of a kind have a higher disposition to exhibit kind-typical properties in contrast to unclassified members. The labelling theory describes how labels cause the targeted people to adjust their behaviour and self-images to conform to the labels, although it does not describe how this can reinforce the theoretical beliefs associated with the labels. Becker's (1953, 1963) account of the labelling theory, for example, is based primarily on congruent mechanisms mediated by intentional pathways. He argues that labels stick when individuals learn and internalize the concepts and meanings associated with them. Scheff's (1966) account of the labelling theory, instead, concentrates more on the influence of societal reactions and material environments. However, a problem with Scheff's approach is that when the inflicted individuals do not find their labels meaningful and natural, they may oppose them. As an example, McLorg and Taub (1987) demonstrate that while anorectics tend to vigorously and openly deny their label, bulimics find their label more meaningful and natural. The reason is that dieting is not as readily conceived as deviant as excessive eating and vomiting. Moreover, some critics (e.g. Gibbs 1971) of the labelling theory have argued that labels cannot be the initial or primary cause of deviant acts because also unlabelled individuals in similar situations perform them. The explanatory domain of the labelling theory is limited especially in cases where there is an underlying psychiatric disorder (cf. Gove 1975). Nevertheless, this does not rule out the fact that explanations based on labels and feedback effects can supplement other explanations. According to Link et al.'s (1989) modified labelling theory, for instance, limited social opportunities together with internalized expectations of being socially rejected, may reinforce patterns of behaviour and conditions that have resulted from other causes. In criminology, for example, labelling effects mediated by intentional and structural mechanisms are used to predict the development of and propensity for criminal behaviour.

Congruent feedback mechanisms can supplement the explanatory domain of psychiatric disorders in several ways. This is illustrated by the already mentioned anthropological study led by Luhrmann et al. (2015), which indicates that schizophrenia voice hallucinations in the USA are harsh in comparison to Ghana

and India, where they can be guiding and playful. The study and a book by Luhrmann and Marrow (2016), seem to suggest that one of the reasons for the voice-hearing experiences in the USA is that the prototypical expectations associated with the diagnostic category bring about negative social consequences.[9] In the West, the prototype of schizophrenia holds that it is chronic and devastating. In India and Ghana, on the other hand, patients and their families rarely know or remember the diagnosis, but instead interpret the disorder's symptoms in culturally meaningful ways. This means that the individuals suffering from the symptoms do not expect, or are not expected by others, to become failures in social and professional life (p. 205). The stigmatizing conceptions associated with the diagnostic label may also indirectly influence the severity and outcome of schizophrenia. According to two major longitudinal studies of over 30 years conducted by the WHO, people who had received a schizophrenia diagnosis, for instance, in India, Nigeria and Columbia, suffered a milder form of the disorder than people with the diagnosis in the USA, Denmark and Taiwan (Hopper, Jenkins, and Barret 2004). Approximately 50 per cent of the people diagnosed with schizophrenia are less impaired in the global south than in the developed world (Hopper et al. 2007). The dominant biomedical explanations of schizophrenia consider the Western feedback effects as stable background conditions and are therefore sensitive to the symptomatic, severity and outcome variations of the disorder.[10] If the feedback explanation is integrated to supplement the explanation's applicability domain, it becomes insensitive to these variations, while the explanatory scope is widened to enable more precise comparisons between different cultural, historical and individual manifestations of the disorder.

Incongruent mechanisms describe how feedbacks generate, reinforce and maintain destabilizing loops between classifications and interactive kinds. Intentionally mediated incongruence describes how a classification is found meaningless, unpreferred or unnatural by the classified persons, and therefore causes them individually or concertedly to act against it (see Hacking 2007a). As an example, misclassifications can cause incongruence because they are found unfitting and meaningless. Similarly, incongruence can be due to a mismatch between a classificatory conception and the classified people's perceived self-images, experiences and behaviour. On the other hand, structurally mediated incongruence describes how a new classification undermines the practices and structures that brought about the behaviour in the first place, leading to alterations in the classified people's behaviours (Hacking 1995a, ch. 4; Mallon 2016, p. 172).

---

9 See also Kent and Wahass (1996), Mawson et al. (2011), Woods et al. (2014).

10 Pöyhönen (2010, 2013) makes a similar argument about norm dependency of bulimia.

That is, as classified individuals succumb under changed structural constraints and opportunities, they may prefer, find meaningful and natural novel behaviours.

The cross-cultural application of the diagnostic categories in the DSM-5 and the ICD-10 psychiatric classification manuals causes incongruence. The categories are primarily based on symptom delineations drawn from Western patients and informed by Western folk psychology (e.g. individualistic view of the self) (cf. La Roche, Fuentes, and Hinton 2015).[11] Consequently, individuals with different cultural backgrounds may be misdiagnosed, while correct diagnoses can induce opposition because they are found stigmatizing by some cultural minorities (cf. Kirmayer 2001). Both situations may lead to alterations in the behaviour of the diagnosed individuals and thereby further distort theoretical beliefs about the disorders. As part of "modernization" processes, the diagnostic categories can bring about destabilizing cultural and social tensions by transforming local structures and conceptions.[12] By this I mean that the categories may distort or alter culturally different ways of experiencing, manifesting and coping with psychiatric problems (see Kleinman 1988; see Kirmayer 2002). As an example, Kitanaka (2012) argues that the moods labelled as depression in the West were not commonly pathologized in Japan until the introduction of the diagnostic category at the end of the 1990s. The diagnosis and the use of antidepressants have induced unpredictable alterations in patients, leading to further changes in the conception of depression (p. 184). A diagnostic category can also become part of a causal process that enables a novel condition. As an example, the conception of self-starvation as a sign of personal suffering (and the idolization of slimness) interacts with local cultures in Asia, creating dynamic and evolving forms of anorexia nervosa (Lee 1996; Pike and Dunne 2015). Finally, research in cognitive science and psychology indicates that even underlying cognitive mechanisms of disorders may be socially and culturally shaped (Murphy 2015; Washington 2016). In the light of these examples, cross-cultural research can supplement explanations of psychiatric disorders by undermining some of the taken-for-grantedness of Western diagnostic feedback effects.

Incongruent feedback loops that describe the influence of non-epistemic classificatory adjustments arguably represent the biggest obstacle for human kinds to support epistemic projects (cf. Griffiths 1997; cf. Khalidi 2013). As an example, the value-laden shifts in the conceptions associated with ADHD have influenced those classified with the condition. The diagnostic category has strong political and moral implications, and therefore motivates individuals as well as

11 Folk-psychological conceptions shape psychiatric disorders, see Luhrmann (2011).

12 Structural changes can also be instigated by endogenous cultural factors, see Sahlins (1985) and Robbins (2004).

larger interest groups to react. These include pressure from ADHD groups, consisting of patients and their families, as well as lobbying from pharmaceutical companies. Moreover, careless use of standardized scales has enabled individuals who lack the symptom profile to pretend to have them, or to become diagnosed mistakenly by the experts or themselves. The upshot is that the distorted picture of ADHD may prompt behaviour reactions from both genuine and misdiagnosed individuals (cf. Allen 2018). Although these changes are hard to explain and anticipate, incongruent feedback explanations can help to identify how the psychiatric disorder is unstable, and thereby contribute to its explanation and classification.

The explanatory domain approach to interactive kinds supports classificatory pluralism (cf. Dupré 1993). A good example is the culture-bound syndrome *latah* found in South East Asia especially among older women in rural areas. The syndrome's symptoms include losing one's self-control when startled by mimicking, cursing and making vulgar gestures. A cultural explanation for the syndrome can be that it is a ritualized role that permits individuals to violate the normal social structure (Lee 1981; cf. Winzeler 1995). This implies that there is a congruent loop between the cultural expectations associated with the role and the pattern of behaviour. On the other hand, Simons (1996) argues that *latah* is a culture-specific variation of a neurocognitive startle-matching syndrome that may include Tourette's and some other culture-specific conditions. In this light, feedback mechanisms could explain the local manifestations of the neurocognitive kind. Moreover, *latah* (or some of its forms) could simultaneously be held as a unique interactive kind upheld by the interaction of the neurocognitive mechanism and the congruent feedback mechanism (cf. Murphy 2006, p. 276). This means that whether a kind concept (and the property cluster it picks out) is split or lumped depends on the underling mechanisms and the explanatory relevance set by the discipline-dependent characterization of the explanandum.[13] Indeed, when classifications have different epistemic aims, they may benefit from carving the property cluster kind differently (i.e. emphasizing different mechanisms), because explanatory dimensions may have trade-offs (cf. Pöyhönen 2014). For instance, clinical practice may benefit from psychiatric disorder concepts associated with deep proximal explanations, whereas epidemiological approaches may prefer disorder concepts associated with explanations with wide scopes (see Campaner 2014, p. 99). In some cases, there may also be a trade-off between shallower explanations with wider time scopes and deeper explanations with limited time scopes.

13 Psychiatric disorders with neurocognitive subtypes can be split or lumped based on the same argument.

Knowledge of the feedback mechanisms can inform treatment and policy decisions by facilitating interventions that would weaken feedback and increase positive congruence. Luhrmann and Marrow (2016, p. 220), for instance, suggest that the diagnosis of schizophrenia should de-emphasize labelling and focus on behaviour rather than inner experience. The upshot is that individuals could find voice-experiences and treatments more positive and meaningful, which would increase positive congruence (see Thomas et al. 2014). Feedback explanations could also be implemented to mitigate or prevent incongruent effects of exported Western diagnostic categories (see Kirmayer 2001). Even value-driven incongruent loops could support some interventions based on models that represent individual motivations (but cf. Steel 2008, p. 158). In short, congruent explanations may facilitate precise domain-relative interventions because the feedback models are relatively modular. On the other hand, although models of incongruent mechanisms are non-modular or weakly modular, they may nonetheless enable preventive interventions. However, disseminating information about classifications, or impeding their dissemination, can by itself influence the classified people. Hence, modifying classifications based on feedback knowledge does not so much describe classifying kinds more accurately, as describing a co-fitting process that transforms both. This means that the conceptual engineering that a feedback explanation enables is tantamount to kind amending.

Lastly, although I have argued that there is no principled epistemic difference between interactive human and biological kinds, there may be a quantitative explanatory difference. When the looping effect works through the direct or indirect intentions of the classified humans, their reactions are more fine-grained than is the case with environmental looping effects of, for instance, disease entities. The reason is that humans can rationalize, interpret and explain their own behaviour in a fine-grained manner. This means that knowledge of the looping effect is either relevant to enable stable explanations of human kinds, or as with value-driven incongruent loops, to understanding why and how such explanations are unstable and limited. This is exemplified by a self-diagnosing psychiatrist who adjusts his or her behaviour to be congruent or opposed to even the smallest diagnostic changes. The dependency could have systematic counterfactual power under the right motives and constraints, or alternatively, it could help to understand why such explanations are highly limited. On the other hand, counterfactual dependencies between classificatory practices and the behaviour of some disease entities are so course-grained, that they provide only weak explanations and projections.

4 Conclusions

I have defended a realistic account of interactive human kinds by arguing that feedback explanations can supplement the explanatory domains over which the kinds sustain epistemic projects. A feedback explanation's explanatory relevance depends on its ability to widen and deepen – enhance the domain of applicability – ways of explaining a human kind. By applying this approach to empirical case studies, I demonstrated that congruent mechanisms can supplement, in various ways and degrees, the applicability domain of interactive human kinds, while incongruent mechanisms can help to identify why the domain is unstable. However, because human categorizations and their effects are constantly reproduced in social interactions, the epistemic projects that human kinds support are in general not as stable as the projects that prototypical natural kinds support. The underlying idea, nevertheless, is that whether an explanation of a human kind is sufficiently stable depends on the discipline-relative epistemic aims set for the explanatory domain.

An important implication of my account is that classificatory and diagnostic practices should pay attention to the mechanisms that underlie the dynamics of interactive kinds such as psychiatric disorders. Feedback explanations may not only contribute to explaining the properties of kinds and predicting when a classificatory adjustment is needed, but also facilitate conceptual engineering and thereby enable kind amendments. Just as criminology is used in assessing criminal policy measures, feedback explanations could be implemented to predict and mitigate the negative effects of diagnostic practices.[14]

References

Allen, S. R. 2018. "Kinds Behaving Badly: Intentional Action and Interactive Kinds." *Synthese* 1–30, https://doi.org/10.1007/s11229-018-1870-0.

Appiah, K. A. 2005. *The Ethics of Identity*. Princeton, NJ: Princeton University Press.

Becker, H. 1953. "Becoming a Marihuana User." *American Journal of Sociology* 59: 235–42.

Becker, H. 1963. *Outsiders: Studies in the Sociology of Deviance*. New York: The Free Press of Glencoe.

Biggs, M. 2009. "Self-Fulfilling Prophecies." In *The Oxford Handbook of Analytical Sociology*, edited by P. Hedström, and P. Bearman, 294–314. Oxford: Oxford University Press.

14 I would like to thank Petri Ylikoski, Gabriel Sandu and Alexander Bird for comments on earlier versions of this paper.

Bird, A., and E. Tobin 2008. (Spring 2018 Edition). "Natural Kind." In *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. https://plato.stanford.edu/archives/spr2018/entries/natural-kinds/ (accessed September 1, 2018).

Boyd, R. 1991. "Realism, Anti-foundationalism and the Enthusiasm for Natural Kinds." *Philosophical Studies* 61: 127–48.

Boyd, R. 1999. "Kinds as the 'Workmanship of Men', Realism, Constructivism, and Natural Kinds." In *Rationalität, Realismus, Revision: Proceedings of the Third International Congress, Gesellschaft für Analytische Philosophie*, edited by J. Nida-Rümelin. Berlin: de Gruyter.

Campaner, R. 2014. "Explanatory Pluralism in Psychiatry: What Are We Pluralists about, and Why?." In *New Directions in the Philosophy of Science, The Philosophy of Science in a European Perspective 5*, edited by M. Galavotti, D. Dieks, W. Gonzalez, S. Hartmann, and T. Uebel, 87–103. Switzerland: Springer International Publishing.

Coleman, J. 1990. *Foundations of Social Theory*. Cambridge, MA: Belknap Press.

Collins, R. 2004. *Interaction Ritual Chains*. Princeton, NJ: Princeton University Press.

Cooper, R. 2004. "Why Hacking is Wrong about Human Kinds." *The British Journal for the Philosophy of Science* 55: 73–85.

Douglas, M. 1986. *How Institutions Think*. Syracuse, NY: Syracuse University Press.

Dupré, J. 1993. *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Cambridge, MA: Harvard University Press.

Ellis, B. 2001. *The Philosophy of Nature: A Guide to the New Essentialism*. Chesham England: Acumen.

Elster, J. 2015. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

First, M. B. 2017. "Factors in the Development of Psychiatric Epidemics." In *Philosophical Issues in Psychiatry IV. International Perspectives in Philosophy and Psychiatry*, edited by K. Kendler, and J. Parnas, 130–42. Oxford: Oxford University Press.

Foucault, M. 1978. "The History of Sexuality." In *An Introduction*, Vol. I. New York: Pantheon.

Geertz, C. 1973. *The Interpretation of Cultures*. New York: Basic Books.

Gibbs, J. P. 1971. "A Critique of the Labeling Perspective." In *The Study of the Social Problems*, edited by E. Rubington, and M. Weinberg, 193–205. Oxford: Oxford University Press.

Gove, W. R. 1975. "Labelling and Mental Illness." In *The Labelling of Deviance: Evaluating a Perspective*, edited by W. Gove, 35–81. New York: Halsted.

Griffiths, P. 1997. *What Emotions Really Are*. Chicago: The University of Chicago Press.

Griffiths, P. 1999. "Squaring the Circle: Natural Kinds with Historical Essences." In *Species. New Interdisciplinary Essays*, edited by R. A. Wilson. Cambridge, MA: MIT Press.

Hacking, I. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

Hacking, I. 1986. "Making Up People." In *Reconstructing Individualism*, edited by T. Heller, M. Sosna, and D. Wellbery, 222–36. Stanford, CA: Stanford University Press.

Hacking, I. 1991. "A Tradition of Natural Kinds." *Philosophical Studies* 61: 109–26.

Hacking, I. 1995a. *Rewriting the Soul: Multiple Personality and the Sciences of Memory*. Princeton: Princeton University Press.

Hacking, I. 1995b. "The Looping Effects of Human Kinds." In *Symposia of the Fyssen Foundation. Causal cognition: A Multidisciplinary Debate*, edited by D. Sperber, D. Premack, and A. J. Premack, 351–94. New York: Clarendon Press.

Hacking, I. 1997. "Taking Bad Arguments Seriously: Ian Hacking on Psychopathology and Social Construction." *London Review of Books* 19: 14–6.

Hacking, I. 1998. *Mad Travelers*. Cambridge, MA: Harvard University Press.
Hacking, I. 1999. *The Social Construction of What?* Cambridge, MA: Harvard University Press.
Hacking, I. 2002. "How "Natural" are "Kinds" of Sexual Orientation?" *Law and Philosophy* 21: 335–47.
Hacking, I. 2007a. "Kinds of People: Moving Targets." *Proceedings of the British Academy*, 151, 285–318.
Hacking, I. 2007b. "Natural Kinds: Rosy Dawn, Scholastic Twilight." *Royal Institute of Philosophy Supplement* 61: 203–39.
Haslanger, S. 2012. *Resisting Reality. Social Construction and Social Critique*. Oxford: Oxford University Press.
Hauswald, R. 2016. "The Ontology of Interactive Kinds." *Journal of Social Ontology* 2: 203–21.
Hedström, P., and P. Ylikoski. 2010. "Causal Mechanisms in the Social Sciences." *Annual Review of Sociology* 36: 49–67.
Henderson, D. K. 1993. *Interpretation and Explanation in the Human Sciences*. Albany: State University of New York.
Hopper, K., J. Jenkins, and R. Barret. 2004. "Interrogating the Meaning of 'Culture' in the WHO International Studies of Schizophrenia." In *Schizophrenia, Culture, and Subjectivity*, 62–87. Cambridge, UK: Cambridge University Press.
Hopper, K., G. Harrison, A. Janca, and N. Sartorius. 2007. *Recovery from Schizophrenia. An International Perspective. A Report from the WHO Collaborative Project, The International Study of Schizophrenia*. Oxford: Oxford University Press.
Kendler, K., P. Zachar, and C. Craver. 2011. "What Kinds of Things are Psychiatric Disorders." *Psychological Medicine* 41: 1143–50.
Kent, G., and S. Wahass. 1996. "The Content and Characteristics of Auditory Hallucinations in Saudi-Arabia and the UK: A Cross-Cultural Comparison." *Acta Psychiatrica Scandinavica* 94 (6): 433–7.
Khalidi, M. A. 2010. "Interactive Kinds." *The British Journal for the Philosophy of Science* 61: 335–60.
Khalidi, M. A. 2013. *Natural Categories and Human Kinds: Classification in the Natural and Social Sciences*. Cambridge, UK: Cambridge University Press.
Kincaid, H. 1996. *Philosophical Foundations of the Social Sciences. Analyzing Controversies in Social Research*. Cambridge: Cambridge University Press.
Kirmayer, L. J. 2001. "Cultural Variations in the Clinical Presentation of Depression and Anxiety: Implications for Diagnosis and Treatment." *Journal of Clinical Psychiatry* 62: 22–30.
Kirmayer, L. J. 2002. "Psychopharmacology in a Globalizing World: The Use of Antidepressants in Japan." *Transcultural Psychiatry* 39: 295–322.
Kitanaka, J. 2012. *Depression in Japan. Psychiatric Cures for a Society in Distress*. Princeton: Princeton University Press.
Kleinman, A. 1988. *Rethinking Psychiatry: From Cultural Category to Personal Experience*. New York: The Free Press.
Kokkonen, T., and I. Koskinen. 2016. "Genres as Real Kinds and Projections. Homeostatic Property Clusters in Folklore and Art." In *Genre – Text – Interpretation. Multidisciplinary Perspectives on Folklore and Beyond. Studia Fennica Folkloristica 22*, edited by K. Koski, Frog, and U. Savolainen, 89–109. Helsinki: Finnish Literature Society, SKS.
Kornblith, H. 1993. *Inductive Inference and Its Natural Ground: An Essay in Naturalistic Epistemology*. Cambridge, MA: The MIT Press.
Kripke, S. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kuorikoski, J., and S. Pöyhönen. 2012. "Looping Kinds and Social Mechanisms." *Sociological Theory* 30: 187–205.
Laimann, J. 2018. "Capricious Kinds." *The British Journal for the Philosophy of Science* 71 (3): 1043–68.
Lakoff, A. 2000. "Adaptive Will: The Evolution of Attention Deficit Disorder." *Journal of the History of the Behavioral Sciences* 36: 149–69, https://doi.org/10.1002/(sici)1520-6696(200021)36:2<149::aid-jhbs3>3.0.co;2-9.
Lee, R. L. M. 1981. "Structure, and Anti-structure in the Culture-Bound Syndromes: The Malay Case." *Culture, Medicine and Psychiatry* 5: 233–48.
Lee, S. 1996. "Reconsidering the Status of Anorexia Nervosa as a Western-Bound Syndrome." *Social Science & Medicine* 42: 21–34.
Link, B. G., F. T. Cullen, E. Struening, P. E. Shrout, and B. P. Dohrenwend. 1989. "A Modified Labeling Theory Approach to Mental Disorders: An Empirical Assessment." *American Sociological Review* 54: 400–23.
Lucas, R. 1976. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets. Carnegie-Rochester Conference Series on Public Policy 1*, edited by K. Brunner, and A. H. Meltzer, 19–46. New York: American Elsevier.
Luhrmann, T. 2011. "Hallucinations and Sensory Overrides." *Annual Review of Anthropology* 40 (1): 71–85.
Luhrmann, T., R. Padmavati, H. Tharoor, and A. Osei. 2015. "Difference in Voice-Hearing Experiences of People with Psychosis in the USA, India and Ghana: Interview-Based Study." *The British Journal of Psychiatry* 206: 41–4.
Luhrmann, T., and J. Marrow, eds. 2016. *Our Most Troubling Madness: Case Studies in Schizophrenia across Cultures*. Oakland, CA: University of California Press.
MacKenzie, D. 2008. *An Engine, Not a Camera: How Financial Models Shape Markets*. Cambridge, MA: The MIT Press.
Mallon, R. 2003. "Social Construction, Social Roles, and Stability." In *Socializing Metaphysics: The Nature of Social Reality*, edited by F. F. Schmitt, 327–54. Lanham, MD: Rowman & Littlefield.
Mallon, R. 2016. *The Construction of Human Kinds*. Oxford: Oxford University Press.
Marchionni, C. 2008. "Explanatory Pluralism and Complementarity: From Autonomy to Integration." *Philosophy of the Social Sciences* 38: 314–33.
Martínez, M. L. 2009. "Ian Hacking's Proposal for the Distinction between Natural and Social Sciences." *Philosophy of the Social Sciences* 2: 221–34.
Mawson, A., K. Berry, C. Murray, and M. Hayward. 2011. "Voice Hearing within the Context of Hearers' Social Worlds: An Interpretative Phenomenological Analysis." *Psychology and Psychotherapy: Theory, Research and Practice* 84: 256–72.
McLorg, P. A., and D. E. Taub. 1987. "Anorexia Nervosa and Bulimia: The Development of Deviant Identities." *Deviant Behavior* 8: 177–89.
Millikan, R. 1999. "Historical Kinds and the "Special Sciences"." *Philosophical Studies* 95: 45–65, https://doi.org/10.1023/a:1004532016219.
Murphy, D. 2001. "Hacking's Reconciliation: Putting the Biological and Sociological Together in the Explanation of Mental Illness." *Philosophy of the Social Sciences* 31: 139–61.
Murphy, D. 2006. *Psychiatry in the Scientific Image*. Cambridge, MA: The MIT Press.
Murphy, D. 2015. "'Deviant Deviance': Cultural Diversity in DSM-5." In *The DSM-5 in Perspective: Philosophical Reflections on the Psychiatric Babel*, edited by S. Demazeux, and P. Singy, 97–110. Dordrecht: Springer.

Pike, K. M., and P. E. Dunne. 2015. "The Rise of Eating Disorders in Asia: A Review." *Journal of Eating Disorders* 3 (33), https://doi.org/10.1186s40337-015-0070-2.
Putnam, H. 1975. "The Meaning of Meaning." In *Mind, Language and Philosophy – Philosophical Papers*, Vol. 2, 215–71. Cambridge: Cambridge University Press.
Pöyhönen, S. 2010. Natural Kinds with Extended Mechanisms, Rough Draft. https://www.ed.ac.uk/files/atoms/files/ppig_extended_mechanisms.pdf (accessed October 1, 2018).
Pöyhönen, S. 2013. "Carving the Mind by its Joints: Culture-Bound Psychiatric Disorders as Natural Kinds." In *Regarding the Mind, Naturally: Naturalist Approaches to the Science of the Mental*, edited by K. Talmont-Kaminsky, and M. Milkowski, 30–48. Newcastle upon Tyne: Cambridge Scholars Publishing.
Pöyhönen, S. 2014. "Explanatory Power of Extended Cognition." *Philosophical Psychology* 27: 735–59.
La Roche, M., M. Fuentes, and D. Hinton. 2015. "A Cultural Examination of the DSM-5: Research and Clinical Implications for Cultural Minorities." *Professional Psychology: Research and Practice* 46 (3): 183–9.
Reydon, T. A. C. 2009. "How to Fix Kind Membership: A Problem for HPC Theory and a Solution." *Philosophy of Science* 76: 724–36.
Robbins, J. 2004. *Becoming Sinners – Christianity and Moral Torment in a Papua New Guinea Society*. Berkeley: University of California Press.
Rosenberg, A. 2016. *Philosophy of Social Science*. Boulder, CO: Westview Press.
Root, M. 2000. "How We Divide the World." *Philosophy of Science* 67: 628–39.
Sahlins, M. 1985. *Islands of History*. Chicago: University of Chicago Press.
Scheff, T. J. 1966. *Being Mentally Ill*. Chicago: Aldine.
Searle, J. R. 1996. *The Construction of Social Reality*. London: Penguin.
Siegel, J. 1997. *Fetish, Recognition, Revolution*. Princeton: Princeton University Press.
Simons, R. 1996. *Boo! Culture, Experience, and the Startle Reflex*. Oxford: Oxford University Press.
Steel, D. 2006. "Methodological Individualism, Explanation, and Invariance." *Philosophy of the Social Sciences* 36: 440–63.
Steel, D. 2008. *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford: Oxford University Press.
Taylor, C. 1971. "Interpretation and the Sciences of Man." *The Review of Metaphysics* 25: 1–51.
Thomas, N., M. Hayward, E. Peters, M. van der Gaag, R. P. Bentall, J. Jenner, C. Strauss, I. E. Sommer, L. C. Johns, F. Varese, J. M. García-Montes, F. Waters, G. Dodgson, and S. McCarthy-Jones. 2014. "Psychological Therapies for Auditory Hallucinations (Voices): Current Status and Key Directions for Future Research." *Schizophrenia Bulletin* 40 (Suppl. 4): 202–12.
Tuomela, R. 1977. *Human Action and Its Explanation: A Study on the Philosophical Foundations of Psychology*. Dordrecht: Reidel.
Washington, N. 2016. "Culturally Unbound: Cross-Cultural Cognitive Diversity and the Science of Psychopathology." *Philosophy, Psychiatry, and Psychology* 23 (2): 165–79.
Winch, P. 1958. *The Idea of Social Science*. London: Routledge & Kegan Paul Ltd.
Winzeler, R. L. 1995. *Latah in Southeast Asia: The History and Ethnography of a Culture-Bound Syndrome*. Cambridge: Cambridge University Press.
Woods, A., N. Jones, M. Bernini, F. Callard, B. Alderson-Day, J. C. Badcock, V. Bell, C. C. Cook, T. Csordas, C. Humpston, J. Krueger, F. Larøi, S. McCarthy-Jones, P. Moseley, H. Powell, A. Raballo, D. Smailes, and C. Fernyhough. 2014. "Interdisciplinary Approaches to the

Phenomenology of Auditory Verbal Hallucinations." *Schizophrenia Bulletin* 40 (Suppl. 4): 246–54.

Woodward, J. 2000. "Explanation and Invariance in the Special Sciences." *The British Journal for the Philosophy of Science* 5: 197–254.

Woodward, J. 2002. "What is a Mechanism? A Counterfactual Account." *Philosophy of Science* 69: 366–77.

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, J. 2015. "Cause and Explanation in Psychiatry: An Interventionist Perspective." In *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*, edited by K. S. Kendler, and J. Parnas, 132–95. Baltimore, MD: John Hopkins University Press.

von Wright, G. H. 1971. *Explanation and Understanding*. London: Routledge & Kegan Paul Ltd.

Ylikoski, P. 2001. *Understanding Interests and Causal Explanation*. Ph.D. thesis. Department of Social and Moral Philosophy, University of Helsinki. https://helda.helsinki.fi/bitstream/handle/10138/21811/understa.pdf?sequence= (accessed September 1, 2018).

Ylikoski, P. 2007. "The Idea of Contrastive Explanandum." In *Rethinking Explanation. Boston Studies in the Philosophy of Science*, Vol. 252, edited by J. Persson, and P. Ylikoski, 27–42. Dordrecht: Springer.

Ylikoski, P., and J. Kuorikoski. 2010. "Dissecting Explanatory Power." *Philosophical Studies* 148: 201–19.